CNVVE: Dataset and Benchmark for Classifying Non-verbal Voice Expressions

Ramin Hedeshy[1,2], *Raphael Menges*[1,2], *Steffen Staab*[1,2,3]

[1]University of Stuttgart, Germany
[2]Semanux GmbH, Germany
[3]University of Southampton, UK

ramin.hedeshy@ipvs.uni-stuttgart.de, raphael.menges@ipvs.uni-stuttgart.de,
steffen.staab@ipvs.uni-stuttgart.de

Abstract

Non-verbal voice expressions (NVVEs) have been adopted as a means of human-computer interaction in research studies. However, exploring non-verbal voice-based interactions has been constrained by the limited availability of suitable training data and computational methods for classifying such expressions, leading to a focus on simple binary inputs. We address this issue with a new dataset containing 950 audio samples comprising 6 classes of voice expressions. The data were collected from 42 speakers who donated voice recordings. The classifier was trained on the data using features derived from mel-spectrograms. Furthermore, we studied the effectiveness of data augmentation and improved over the baseline model accuracy significantly with a test accuracy of 96.6% in a 5-fold cross-validation. We have made CNVVE publicly accessible in the hope that it will serve as a benchmark for future research.
Index Terms: non-verbal voice recognition, human-computer interaction, speech impairment, dysarthric speech

1. Introduction

Inclusive speech technologies are gaining increasing attention, particularly in the context of enabling individuals with speech disabilities to communicate effectively and interact with speech recognition systems. A current area of active research is the recognition of dysarthric speech, as evidenced by studies such as [1] and [2]. However, there is an inverse relationship between the degree of impairment and the accuracy of speech recognition, which means that modern automatic speech recognition (ASR) systems are not always a feasible solution for people with severe speech impairments. Therefore, finding alternative ways to facilitate communication for such individuals is of great importance.

Interacting via non-verbal or non-lexical voice expressions, such as humming, can be particularly beneficial for people with severe disabilities who may also suffer from speech disorders. Over 96% of people with speech disorders can produce some form of non-speech voice [3]. These individuals may have difficulty in communicating with common language, and may not be able to use conventional voice commands. Moreover, non-verbal voice interactions put less strain on the vocal cords, as they often involve quieter, more subtle movements and less effort from the user, making it a good option for people who are recovering from a speech-related injury. Non-verbal voice interactions can provide an alternative way for these individuals to express themselves and communicate their needs or preferences. For example, humming could be used to control devices or functions, such as turning on a light or activating a motorized wheelchair [4]. Overall, the use of non-verbal voice interactions can greatly improve the accessibility and usability of technology for people with disabilities and can help to break down barriers to communication and expression.

In this work, we introduce and describe a dataset of Non-Verbal Voice Expressions (NVVEs) recorded from healthy individuals and dysarthric speakers. This work aims at providing a novel resource for future developments of non-verbal voice recognition to be used in computer interactions and assistive technologies. We believe such a dataset is an essential basis for developing a feasible classification model. The final dataset, code for creating a voice recognition system, trained models, and other voice processing tools used in our work are publicly available at `https://github.com/hedeshy/CNVVE`. We hope the availability of these resources will enable new assistive technology projects that better serve the needs of the community.

2. Related work

In order to understand the use of NVVEs input for computer interactions, we investigated their proposed recognition methods and application.

Hawley *et al.* [6] collected voice samples from people with severe dysarthria to create an environmental control system based on hidden Markov models (HMMs). Bilmes *et al.* [7] also uses HMMs for pattern recognition on three continuous vocal characteristics energy, pitch, and vowel quality to create a vocal joystick for people with speech impairments. Compared to models based on convolutional neural networks (CNNs), using conventional HMMs is disadvantaged due to their assumptions of linearity, difficult scalability, and less robustness to noise.

Some related works aim at detecting vocal segregates (fillers like 'um' and 'uh', pauses, and other hesitation phenomena) to address disfluencies in speaking [8, 9]. They suggest machine learning models to automatically detect and remove 'um', 'uh', breaths, laughter, and word repetitions. Other forms of human non-speech voice samples such as sneezing, breathing, and coughing are listed in the ESC dataset Piczak [10]. However, these works have been developed with different objectives and are not suitable for immediate interactions.

Various studies have investigated the potential of NVVEs as a means of providing accessible input mechanisms [11, 7, 12]. Non-verbal voice expressions can be detected and when detected undergo binary classification. In this way, they can be used as on/off signals, comparable to a button [13]. Continuous non-verbal voice expressions, such as humming, can be used for a slider [11], or in combination with eye-tracking for hands-free text entry [5]. Additionally, NVVE can serve as an alternative input modality for wheelchair control [4]. Furthermore, it has shown potential in controlling games [14, 15] and even artistic

Table 1: *Examples of NVVEs and their usage.*

Sound	Usage	HCI usage example
"Uh-huh" or "mm-hmm"	Confirmation, affirmation	Item selection [5] & interaction with Yes/No-prompts
"Uh-uh" or "mm-mm"	Disagreement, negation	Correction & interaction with prompts
"Hush" or "Shh"	Request for quiet	Silent a smartphone
"Psst"	Attention-getting	Wake up a device or a specific app
"Ahem"	Attention-getting, disapproval or embarrassment	Wake up a device or a specific app & disapproval
Continuous humming, "hmmm"	Indication of thinking or considering something	Temporal interactions, e.g., Wheelchair control [4]

expression [16]. Table 1 provides examples of NVVEs and their potential usage in human-computer interaction (HCI).

3. Collecting non-verbal voice expressions

After surveying previous research in HCI, we compiled a list of NVVEs that can be leveraged for computer interaction. While some NVVEs have been reported in scientific publications in the field of HCI, there are others that have not yet been documented but could still be used for discrete and continuous inputs, listed in Table 1. We recorded these voice expressions as well to facilitate the creation of assistive interactive technologies in the future.

3.1. Procedure

We have developed a dedicated website for data collection that defines the purpose and type of voice data we seek to collect by providing example recordings to participants as well as the expressions' written equivalent, e.g., "Uh-huh". The website presents the list of NVVE (Table 1) and a recorder to users so that they can record their voice samples. They have the opportunity to supplement the recording with additional information such as their gender, age, and the ambient environment and noise level at the time of recording. The website is designed for accessibility via smartphone, allowing participants to record their voices using their own devices. Recording through a smartphone microphone accurately captures real-life audio and minimizes the mismatch between the experiment and actual conditions caused by different recording equipment. After recording, participants are provided with their recorded samples, allowing them to re-record if they are not satisfied with the initial recording. Audio recordings were automatically saved in the .wav format and kept anonymous, with a sampling rate of 48 kHz and a bit depth of 32 bits.

3.2. Participants

We received recordings from 42 (19 females, 23 males, with an estimated average age of 29.6, SD = 7.84) participants who anonymously and voluntarily donated their voices. Four of them reported having dysarthria. However, it did not hinder them from performing the non-verbal voice expressions and participating in the data collection program.

Participants provided their informed consent by accepting the data collection terms before undergoing the experimental procedures. The protocol was designed according to the data protection declaration and approved by the local ethical committee of the university.

3.3. Dataset

We recorded the aforementioned NVVEs from a total of 42 speakers, each of whom performed from 1 to 5 recording sessions. The dataset contains 950 voice data samples in total, around 150 files in each NVVEs class. The dataset is accompanied by a metadata file that lists the filename, location, label, and, when provided by speakers, their gender and age information. We reviewed and removed any corrupted recorded files.

4. Method for classifying NVVEs

In order to achieve better results for the task of classifying NVVEs, we introduce a model inspired by existing CNN-based audio processing techniques [17, 18]. An overview of our method is shown in Figure 1.

4.1. Data pre-processing and normalization

We first applied the Google WebRTC voice activity detection (VAD) algorithm [19] on the given audio files to remove noise or silence from the collected voice signals. Next, we performed downsampling of the given signals to 16,000 Hz. A sample rate of 16,000 Hz is commonly used for speech and voice recognition tasks, as it provides a good balance between accuracy and computational efficiency [20, 21]. This is because human voice, including plosive and sibilant fricatives, is typically in the range of 100 Hz to 8,000 Hz, such that 16,000 Hz provides enough samples to capture the majority of speech sounds.

To produce uniform mel-spectograms later on, the audio signal lengths must be normalized. If the sample is shorter than 1 second, it is padded with zeros. The 1-second length was selected as the majority of samples are typically shorter, averaging 0.82 seconds.

4.2. Feature extraction

Mel-frequency cepstral coefficients (MFCCs) and mel-spectrograms are common inputs for audio signal processing tasks such as environmental sound identification, speech recognition, and music genre classification. In recent studies, using mel-spectrograms features has shown state-of-the-art performance [22, 23, 8, 2].

We transform the pre-processed waveforms to mel-spectrograms. The reason for the chosen audio representation will be further explained in Section 5.2.1. The mel-spectrogram has a size of 96×63, where 96 represents the number of mel frequency bands and 63 represents the number of time frames. The time frames correspond to a duration of approximately 1 second, since a window size of 512 samples and a hop size of 256 samples were used to generate the spectrogram.

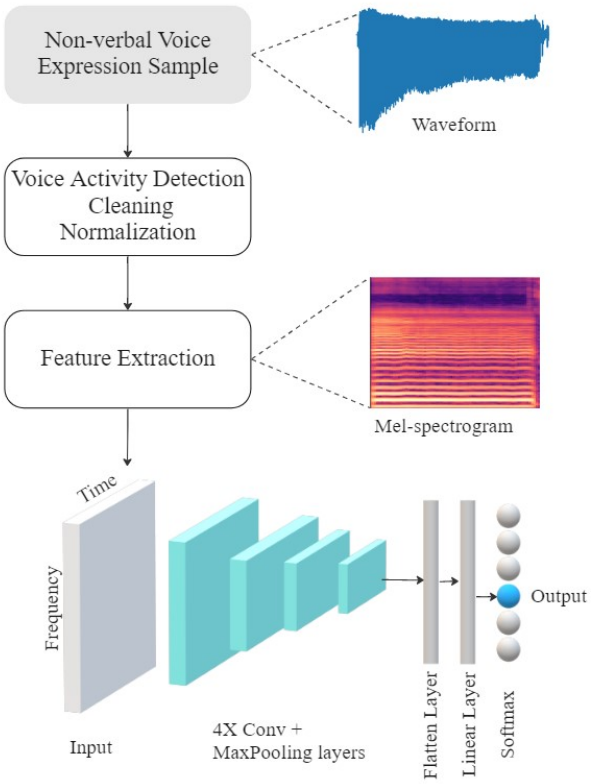

Figure 1: *A block diagram depicting an overview of the system used for non-verbal voice expression classification.*

4.3. A model for classifying non-verbal voice expression

Our network is composed of 4 convolutional layers followed by a flatten layer, with each convolutional layer having a ReLU activation function. The first Conv2D layer receives the input shape with 64 filters, a kernel size of 3, and a stride of 1. We use the 'same padding' operation to ensure that the output of the layer is of the same height and width as the input. This is followed by a 2×2 MaxPooling2D layer to reduce the dimension of the input shape. We repeat this sequence three times with 3×3 convolutions of 128 filters. The output of the convolutional layers is converted to a one-dimensional array by a flatten layer, which is then passed to the next layers with a dropout rate of 0.5. A linear layer is used in the final stages before the output layer with a softmax activation function, which consists of 6 nodes for our classification task. We used ADAM for learning rate control and cross-entropy as the loss function.

We built the model utilizing PyTorch [24] and the audio and speech processing capabilities of the Torchaudio toolkit [25] for data preprocessing and feature extraction.

5. Experiments

5.1. Comparison with the baseline

We compared our systems with a neural-network-based baseline, Piczak-CNN [17], which is designed for classifying environment sounds and which has been used in several studies as baseline [26, 27]. The original Piczak-CNN architecture yielded weak performance in our experiments, so we fine-tuned it (number of mel-bands, learning rate) on CNVVE for a stronger baseline. The models were tested in a cross-validation scheme of 5 folds using an 80/20 training/test split on the dataset. Our model achieves a mean test accuracy of 87.6%, surpassing the accuracy of the Piczak-CNN model, which achieved 82.1%. Figure 2 shows the mean accuracy of our model performance for 6 classes of over 50 epochs.

5.2. Experiments on detection feature sets

5.2.1. MFCC vs. Mel Spectrogram

We observed through experiments that using the mel spectrogram representations for the audio data provided better results compared to Mel-Frequency Cepstral Coefficients (MFCC) representation. We trained the model using both techniques as input with different shape values and the models trained using mel spectrogram on average showed better performance by at least 7%. We note that other existing studies such as [23, 22] also confirm our finding that predictive models based on the mel spectrogram perform better than models based on conventional features, e.g., MFCC.

5.2.2. Finding The Right Number of Mel Bands

The majority of recent speech classification literature based on mel spectrogras suggest mel-spectrograms consisting of at least 64 frequency bands to keep an optimal balance between learning rate and recognition accuracy [23, 8].

By experimenting with mel-spectrograms with 32, 64, 96 and 128 bands as input features, we discovered that the optimal balance between learning rate and recognition accuracy can be reached using mel-spectrograms with 96 bands. Classification test accuracy increased from 76% to 88% after increasing the number of mel bands from 64 to 96. Extending to 128-band mel-spectrograms did not significantly improve the accuracy, however, increased the training time. Using mel-spectrograms consisting of 32 mel-frequency bands lead to a significant increase in error rate.

5.3. Data augmentation

In our data augmentation set of experiments, we trained our network with a set of different augmentation methods on our dataset and evaluated it on unmodified test sets. Figure 3 il-

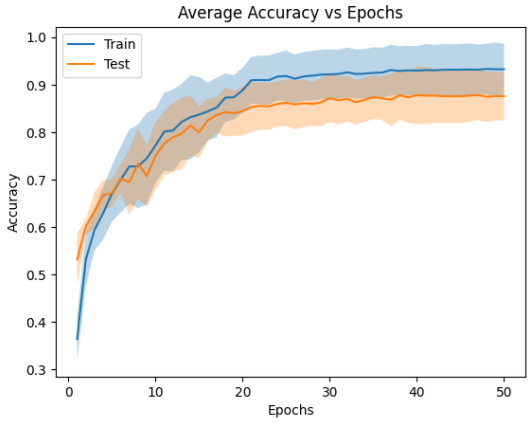

Figure 2: *Standard deviation and mean of the achieved accuracy from 5 folds cross-validation testing.*

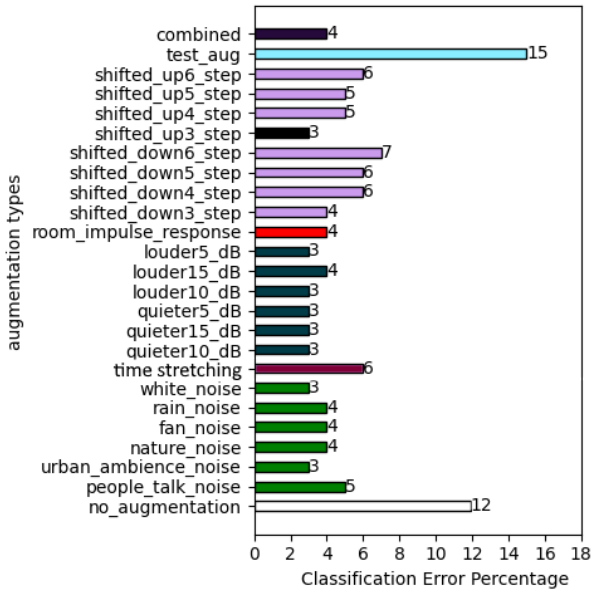

Figure 3: *Classification error for different augmentation methods.*

lustrates the results. The first bar shows the result of the combination of all augmentation methods together and the last bar depicts the base result without any data augmentation. The second bar shows the classification error rate of augmenting the test set, in this case only the test set was augmented leaving the training unmodified. All other lines show the results of a single data augmentation method at a respective strength.

The model accuracy has improved from about 88% to over 96% by applying all the augmentation techniques. Adding background noise effects, such as fan noise, rain noise, conversations in the background, etc., give a modest improvement, with the white noise and urban ambiance noise having the best impact resulting in a significant reduction in classification error of up to 60%. Similarly, loudness change in a range of ±5dB to ±15dB significantly diminishes the error rate by more than a half. Furthermore, applying effects, e.g., room impulse response filter, increased the model accuracy by 4%.

We have also examined the effect of pitch shifting in a range of ±3 to ±6 steps. We chose this range as we noticed no noticeable change in lower values and corruption in some files when increasing the amount further. Shifting the pitch up and down by 3 steps shows the best effect cutting the error rate by half. Shifting the pitch in other values also increased the model accuracy by at least 1%.

After applying the data augmentation to the initial data, the total number of files increased to 41,030.

5.4. Analysis of model performance

Figure 4 shows the confusion matrix of our best-performing model using mel spectogram feature trained on the augmented data. This matrix shows the performance of the model in accurately classifying the test data, with the number of correctly and incorrectly classified samples indicated for each class. We can see that out of 374 unseen testing data, our model has incorrectly classified only 17 testing samples reaching an accuracy of over 95%. Furthermore, it performs well in all the classes

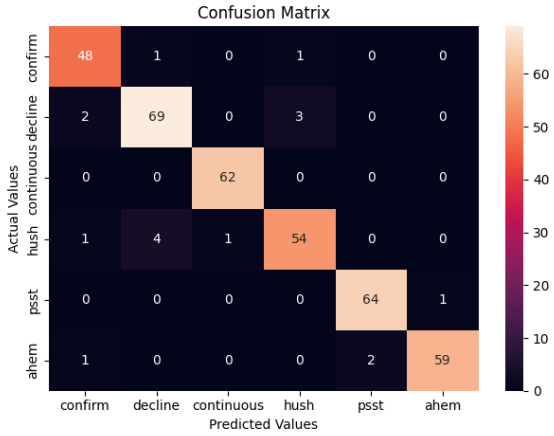

Figure 4: *Confusion matrix of model predictions on unseen testing data.*

with 90% being the least accurate for the class “hush” which is shown in Fig. 4.

6. Conclusions

In this work, we presented CNVVE, a dataset of non-verbal voice expressions, collected from 42 participants, and a CNN-based model that can classify them with high accuracy.

We hope the CNVVE dataset, our proposed data augmentation, classification pipeline, and our experimental results can help the HCI and voice-related research communities and serve as a basis and benchmark for future research.

We trained the model using all the data for training creating a production-ready model that is traced and can be loaded also via the C++ API of PyTorch to be used on various platforms like smart devices with ease. We envision further work to utilize our proposed non-verbal voice input recognition system to create exciting applications in the domain of accessibility, entertainment, and communication.

Additionally, the dataset presented here has the potential to be used for pre-training a network for a different purpose. ASR systems often struggle with transcribing non-lexical fillers, such as “mm-hmm” or “uh-uh,” that occur frequently in spontaneous speech. This can result in the omission of critical information, as these simple expressions can carry significant meaning and express the speaker’s opinions during a conversation.

7. Acknowledgements

We would like to express our sincere gratitude to all the participants who generously donated their voice recordings to our dataset, as well as the technical contribution of Srinivas Kumar to this project. We also extend our thanks to our institutions for their support. This research was partially funded by BMWK/ESF[1] and BMBF[2] under grant no. 03EFRBW231 and 16DHBKI041, respectively.

8. References

[1] K. T. Mengistu and F. Rudzicz, “Adapting acoustic and lexical models to dysarthric speech,” in *2011 IEEE International Con-*

[1] https://https://www.bmwk.de
[2] https://www.bmbf.de

ference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2011, pp. 4924–4927.

[2] Z. Chen, B. Ramabhadran, F. Biadsy, X. Zhang, Y. Chen, L. Jiang, F. Chu, R. Doshi, and P. J. Moreno, "Conformer Parrotron: A Faster and Stronger End-to-End Speech Conversion and Recognition Model for Atypical Speech," in *Proc. Interspeech 2021*, 2021, pp. 4828–4832.

[3] J. McCormack, S. McLeod, L. J. Harrison, and L. McAllister, "The impact of speech impairment in early childhood: Investigating parents' and speech-language pathologists' perspectives using the icf-cy," *Journal of Communication Disorders*, vol. 43, no. 5, pp. 378–396, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021992410000365

[4] N. Peixoto, H. G. Nik, and H. Charkhkar, "Voice controlled wheelchairs: Fine control by humming," *Computer methods and programs in biomedicine*, vol. 112, no. 1, pp. 156–165, 2013.

[5] R. Hedeshy, C. Kumar, R. Menges, and S. Staab, "Hummer: Text entry by gaze and hum," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21. New York, NY, USA: Association for Computing Machinery, 2021. [Online]. Available: https://doi.org/10.1145/3411764.3445501

[6] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria," *Medical Engineering and Physics*, vol. 29, no. 5, pp. 586–593, 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S135045330600138X

[7] J. A. Bilmes, X. Li, J. Malkin, K. Kilanski, R. Wright, K. Kirchhoff, A. Subramanya, S. Harada, J. A. Landay, P. Dowden, and H. Chizeck, "The vocal joystick: A voice-based human-computer interface for individuals with motor impairments," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, p. 995–1002. [Online]. Available: https://doi.org/10.3115/1220575.1220700

[8] G. Zhu, J.-P. Caceres, and J. Salamon, "Filler Word Detection and Classification: A Dataset and Benchmark," in *Proc. Interspeech 2022*, 2022, pp. 3769–3773.

[9] T. Kourkounakis, A. Hajavi, and A. Etemad, "Detecting multiple speech disfluencies using a deep residual network with bidirectional long short-term memory," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6089–6093.

[10] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 1015–1018. [Online]. Available: https://doi.org/10.1145/2733373.2806390

[11] M. Funk, V. Tobisch, and A. Emfield, "Non-verbal auditory input for controlling binary, discrete, and continuous input in automotive user interfaces," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–13. [Online]. Available: https://doi.org/10.1145/3313831.3376816

[12] S. Harada, J. O. Wobbrock, and J. A. Landay, "Voice games: Investigation into the use of non-speech voice input for making computer games more accessible," in *Human-Computer Interaction – INTERACT 2011*, P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, and M. Winckler, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 11–29.

[13] T. Igarashi and J. F. Hughes, "Voice as sound: Using non-verbal voice input for interactive control," in *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 155–156. [Online]. Available: https://doi.org/10.1145/502348.502372

[14] A. J. Sporka, S. H. Kurniawan, M. Mahmud, and P. Slavík, "Non-speech input and speech recognition for real-time control of computer games," in *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, 2006, pp. 213–220.

[15] R. Hedeshy, C. Kumar, M. Lauer, and S. Staab, "All birds must fly: The experience of multimodal hands-free gaming with gaze and nonverbal voice synchronization," in *Proceedings of the 2022 International Conference on Multimodal Interaction*, ser. ICMI '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 278–287. [Online]. Available: https://doi.org/10.1145/3536221.3556593

[16] S. Harada, J. O. Wobbrock, and J. A. Landay, "Voicedraw: A hands-free voice-driven drawing application for people with motor impairments," in *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*, ser. Assets '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 27–34. [Online]. Available: https://doi.org/10.1145/1296843.1296850

[17] k. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2015, pp. 1–6.

[18] H. Zhang, I. Mcloughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 559–563, 2015.

[19] WebRTC, "Webrtc voice activity detection," 2023. [Online]. Available: https://webrtc.org

[20] V. Digalakis, S. Rouvas, and N. Fakotakis, "A comparison of feature extraction techniques for automatic speech recognition," *Speech Communication*, vol. 13, no. 1, pp. 1–14, 1993.

[21] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[22] S. Rao, V. Narayanaswamy, M. Esposito, J. Thiagarajan, and A. Spanias, "Deep learning with hyper-parameter tuning for covid-19 cough detection," in *2021 12th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 2021, pp. 1–5.

[23] M. Lesnichaia, V. Mikhailava, N. Bogach, I. Lezhenin, J. Blake, and E. Pyshkin, "Classification of Accented English Using CNN Model Trained on Amplitude Mel-Spectrograms," in *Proc. Interspeech 2022*, 2022, pp. 3669–3673.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[25] Y.-Y. Yang, M. Hira, Z. Ni, A. Chourdia, A. Astafurov, C. Chen, C.-F. Yeh, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang, J. Lian, J. Mahadeokar, J. Hwang, J. Chen, P. Goldsborough, P. Roy, S. Narenthiran, S. Watanabe, S. Chintala, V. Quenneville-Bélair, and Y. Shi, "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.

[26] J. Sharma, O.-C. Granmo, and M. Goodwin, "Environment Sound Classification Using Multiple Feature Channels and Attention Based Deep Convolutional Neural Network," in *Proc. Interspeech 2020*, 2020, pp. 1186–1190.

[27] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," *IEEE Access*, vol. 7, pp. 7717–7727, 2019.